The disciplinary power of predictive algorithms

de Laat, Paul B.

Published in:
Ethics and Information Technology

DOI:
10.1007/s10676-019-09509-y

Ethics and Information Technology (2019) 21:319–329
https://doi.org/10.1007/s10676-019-09509-y

ORIGINAL PAPER

The disciplinary power of predictive algorithms: a Foucauldian perspective

Paul B. de Laat[1]

Published online: 23 July 2019

Abstract
Big Data are increasingly used in machine learning in order to create predictive models. How are predictive practices that use such models to be situated? In the field of surveillance studies many of its practitioners assert that "governance by discipline" has given way to "governance by risk". The individual is dissolved into his/her constituent data and no longer addressed. I argue that, on the contrary, in most of the contexts where predictive modelling is used, it constitutes Foucauldian discipline. Compliance to a norm occupies centre stage; suspected deviants are subjected to close attention—as the precursor of possible sanctions. The predictive modelling involved uses personal data from both the focal institution and elsewhere ("Polypanopticon"). As a result, the individual re-emerges as the focus of scrutiny. Subsequently, small excursions into Foucauldian texts discuss his discourses on the creation of the "delinquent", and on the governmental approach to smallpox epidemics. It is shown that his insights only mildly resemble prediction as based on machine learning; several conceptual steps had to be taken for modern machine learning to evolve. Finally, the options available to those subjected to predictive disciplining are discussed: to what extent can they comply, question, or resist? Through a discussion of the concepts of transparency and "gaming the system" I conclude that our predicament is gloomy, in a Kafkaesque fashion.

Keywords Algorithms · Discipline · Foucault · Machine learning · Normation · Panopticon · Prediction · Risk · Transparency

Introduction

In our Big Data society, data about persons and their behaviours are collected as a matter of routine. This modern day surveillance—the systematic monitoring of members of society—pervades all areas of life, from work and communication, to consumption and leisure (Lyon 2003). These data are used for purposes of description, pattern recognition, playing various games, and the like. Following a recent development, they can be used for *predictive* purposes. Based on machine learning (henceforth: ML), algorithms are developed that are intended to predict some future behaviour of the persons under consideration. The outcomes assist the managers and authorities involved in taking the best decision. Think of banks granting a loan, insurance companies judging customer claims, tax officials scrutinizing declarations for possible tax evasion, security personnel screening passengers, police officers seeking to prevent crime, etc.

In this article, the focus is on these predictive models. What is the status of these models as institutions apply them? How are their "predictive practices" to be situated? In the field of surveillance studies a theoretical trend can be observed asserting that "governance by discipline" has given way to "governance by risk". Surveillance practices loosen their disciplinary hold on individuals to make way for management of group risks. In this article I take aim at this contention. I argue that the claim is overblown, and that predictive practices in particular still constitute governance by discipline in a wide range of contexts. For this purpose, I take recourse to several texts by Michel Foucault. To his *Discipline and Punish*, obviously, his seminal work on discipline; but also to some of his lectures at the Collège de France. In doing so, I just take out of the Foucauldian tool box the tools that suit me—an approach justly advocated by Kevin Haggerty (2006). In the end, after having made my case for prediction as Foucauldian disciplining, I ask myself

✉ Paul B. de Laat
p.b.de.laat@cerug.nl

[1] University of Groningen, Groningen, The Netherlands

what possibilities remain for subjects involved to comply, question, or resist such modelling and their outcomes? What modes of individual or collective resistance are possible? I argue that our predicament is Kafkaesque, without much prospect of improvement in the near future.

Before commencing the discussion proper, I put some building blocks in place. For one thing, some explications of ML, the basic technique underlying prediction, are in order. For another, I devote a section to *Discipline and Punish*, the "Urtext" as far as societal disciplining is concerned. Thereafter, I come to the core of this article: are predictions enlisted for purposes of governance by discipline or by risk?

Machine learning

First a few words about predictive models based on ML (what follows is based on de Laat 2018). Techniques employed by ML are classification and decision trees, support vector machines, ensemble methods, neural networks, and the like. In inductive fashion an appropriate model is constructed that best fits the data ("training set"). Based on these data, the model is trained step by step, its prediction error ever diminishing. This error is usually expressed through the measures of "precision" (the number of true positives as a fraction of all predicted positives) and "recall" (the number of true positives as a fraction of all positives in the training set)—often combined into the F-measure.[1] In general, the numbers of false positives and false negatives are inversely related: one cannot decrease any one of them without increasing another one.

In the process of modelling, a dominant concern is "overfitting": one goes on and on to train (say) the classifier until the very end. The end product surely fits the training data—but only those; it is unfit to be generalized to other, new data. One recipe against overfitting (amongst many) is to divide the training data into a training set (80%) and a test set (20%). The classifier is trained on the former set, its error diminishing with every iteration. Simultaneously, one keeps an eye on the classifier's error as applied to the other test set. When the latter error starts to increase, it is time to stop and be satisfied with the classifier a few steps back ("early stopping").

A further problem that needs to be taken into account is the "class imbalance problem". In many areas, the "class variables" of the "target variable" (i.e., the subcategories of the main variable to be predicted) are represented very unequally in the population. Think of transactions that amount to monetary fraud, or tax evasion—these only make up a tiny fraction of all transactions. Or think of the fraction of ill-intentioned plane passengers amongst all passengers. Training on such an imbalanced dataset may produce a model that overfits to the majority of data representing bona-fide transactions. The main approach is to adjust the available training set in order to obtain a more balanced set. Either one deletes data points from the overrepresented class ("undersampling"), or adds data points from the underrepresented class ("oversampling").

Note, finally, that the algorithmic end-products of ML are often difficult to interpret, even by experts. The algorithms yield accurate outcomes, but an explanation in understandable terms of why a specific decision is recommended cannot be supplied. The model is effectively a black box for all of us, laymen and experts alike. This is called the problem of "interpretability" or "explainability".

Discipline and Punish

Before tackling the issue whether Foucauldian discipline has withered or not, let me briefly rehearse some arguments as developed by Michel Foucault in *Discipline and Punish* (1977; French original 1975). He analysed the development of the prison regime from the seventeenth century onwards. Discipline enters the prisons, and subsequently the army, schools, hospitals, and factories. Their subjects are variably disciplined by means of a division of tasks, regulations, time-tables, exercises, inspections, examinations, and the like. Furthermore, sanctions and rewards become part and parcel of the disciplinary regime. Foucault stresses that these modalities of power did not overturn the institutions, but just crept into and strengthened them (Foucault 1977, pp. 215, 216).

The disciplinary gaze is exercised along the *panoptic principle* (Foucault 1977, pp. 195–230): one is always in full light and visible, and must assume one is watched all the time. The archetypical design for prisons along this principle had been sketched by Bentham: the Panopticon. A central guard can observe all the inmates all the time; these can neither see each other nor whether the guard is actually observing them. One must be careful though not to equate the panoptic design with Bentham's specific proposal: the principle can have many other applications, in other guises, in hospitals, schools, and factories. Panopticism is a utopian idea; 'in fact a figure of political technology that may and must be detached from any specific use' (Foucault 1977, p. 205).

All the observations gathered serve a purpose: classifying the subjects on a homogeneous scale (or on a number of them). The observance of rules and regulations, the results of examinations and exercises, the displayed morality of behaviour: all of this can be taken into account. The scale not only describes; it *normalizes* at the same time.

[1] $F = 2\,(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$.

In *Discipline and Punish* this process of normation has the following meaning. All subjects involved are drawn into a comparative field which, by its very existence, exerts a pressure to conform. The scale is constructed 'to function as a minimal threshold, as an average to be respected or as an optimum towards which one must move' (Foucault 1977, p. 183). As Foucault summarizes the process: the scale 'compares, differentiates, hierarchies, homogenizes, excludes. In short, it *normalizes'* [italics in original] (Foucault 1977, p. 183). This normalizing power had invaded all the institutions that he examined.

From discipline to risk?

It is this model of panoptic discipline that is declared to be outdated by many scholars in current surveillance studies. Consider David Lyon, a prominent scholar in the field. According to him, the metaphors of Big Brother and the Panopticon have become 'less relevant' (Lyon 2003, p. 19). In some ways, modern day surveillance can be interpreted as 'positive and beneficial, permitting new levels of efficiency, productivity, convenience, and comfort' (Lyon 2003, p. 19). This is achieved by the creation of profiles and risk categories. These are the basis for a process of "social sorting": 'classifying people and populations according to varying criteria, to determine who should be targeted for special treatment, suspicion, eligibility, inclusion, access, and so on' (Lyon 2003, p. 20). Such sorting thus affects people's lifestyle choices and life-chances. This topos of "social sorting" as expressed by Lyon is still very much alive; cf. for example Erwin (2015, pp. 41, 42) who, in a discussion of post-modern "smart surveillance", argues along the very same lines. Note, though, that in a later anthology Lyon seems to retract from this position and blow new life into the panoptic metaphor: 'The Panopticon refuses to go away', and 'the idea of the Panopticon still appears routinely in surveillance discourses' (Lyon 2006, p. 4).

So, with (the earlier) Lyon we see classical discipline moving into the background, broader governance taking over, and associated with this, a shift of focus from the individual to the group. Such theoretical tendencies are even more pronounced with other leading scholars of surveillance, most of them having affinities with the French philosophical tradition. Let me treat some of them.

Early on, Mariana Valverde and Michael Mopas, working in the field of "critical" criminology, already signalled a shift from governance by discipline to governance by risk. In former days, they argue, discipline governed individuals on an individual basis (à la Foucault). Nowadays, the individual is broken up into measurable risk factors (Valverde and Mopas 2004, p. 240). Henceforth governance targets groups identified as high risk in appropriate ways ("targeted governance"). The emphasis on risk groups implies that no efforts are taken any longer to separate the deviant from the normal, the criminal from the honest citizen (Valverde and Mopas 2004, p. 243).

Louise Amoore, also occupied with problems of security, subscribes to this analysis. ML is based on an ontology of association between disparate data (Amoore 2011, p. 27). These are obtained by dissolving the individual into his/her data as deemed appropriate. From the exercise a risk score or flag is derived for each individual (Amoore 2011, p. 25). This data derivative 'is not centred on who we are, nor even on what our data says about us, but on what can be imagined and inferred about who we might be (…)' (Amoore 2011, p. 28). Thus, the future is drawn into the present and action can be taken—the data have become "actionable" (Amoore 2011, p. 29). Predictive algorithms hold the promise that henceforth, risks can be managed ahead of time (Amoore 2013, p. 9).

With Antoinette Rouvroy, the disappearance of the individual becomes ever more pronounced. Regulation by law is gradually replaced by regulation by algorithms. Such algorithmic governance no longer addresses the subject as a moral agent (Rouvroy 2012, p. 11). Instead, it dissolves the individual into a bundle of data, needed for the production of appropriate profiles ("data behaviourism"). Embodied individuals are ignored; only the conditions of their environment are adapted in accordance with algorithmic outcomes (Rouvroy 2012, p. 11). Again, as with Valverde, Mopas, and Amoore, the management of risks and opportunities takes centre stage.

Of course, the aforementioned authors are all tributary to Gilles Deleuze, especially to his "post-scriptum" about the societies of control (Deleuze 1992; original in French 1990). In this short but provocative (as well as wide ranging and speculative) essay he announces the decay of the disciplinary society and the coming of the society of control. Closed disciplinary spaces are replaced by the open spaces (networks) of the control society. Fixed moulds make way for modulating spaces; society's members are constantly circulating. In such a society, individuals are no longer important; they are henceforth reduced to mere "dividuals", available for being divided into their constituent data. These data are processed by computers, which only know the logic of samples, categories, and markets.

Prediction as Foucauldian discipline

The basic tenet that runs through all these accounts is, that disciplining has receded into the background and people are no longer addressed as individuals (which discipline and law have always done). Governance by disciplining individuals has given way to governance by managing the risks that

specific groups represent. Predictive modelling is considered to be the pinnacle of this trend. While some elements of this theorizing may have validity, I think that basically these theories lean on outdated conceptions of what profiling stands for. At the present day, the individual has re-emerged as the focus of attention. Governance by group risk is still adhered to (for example, in police circles), but governance by *individualized* risk is coming to the fore. Focusing on the individual marks its re-entry from the shadows it was deemed to have been relegated to. And let us not be mistaken: such "attention" is nothing other than a disciplinary gaze.

I begin my argumentation with a short overview of the societal sectors where predictive modelling is being used—which, obviously, is just meant to be indicative of which sectors are involved, not exhaustive or representative in any statistical sense. This is a broad spectrum, from private to public organizations. It is in the private sector that data mining took its first steps long ago. Companies perform extensive data collection and modelling, in order to sort their customers into various categories; subsequently, these can be targeted in optimal fashion ("data base marketing"). Soon enough other organizations followed suit. Banks, insurance companies, and tax departments rely on profiling extensively, in order to deal with financial malperformance or outright fraud. Moreover, security forces in general (such as the police, border and airport officials) use it par excellence to combat crime, public disturbances, terrorism and the like. Local authorities, for their part, have more recently initiated innovative prediction efforts in order to curb youth crime, child abuse and domestic violence, and fraud with social security benefits or local taxes. Finally, we must not forget the classical sectors as analysed by Foucault (some public, some private): prisons, the army, schools, hospitals, and factories. Recently, predictive modelling has been gaining a foothold in those sectors as well. Take education: some schools engage in predicting the future performance of their teachers, while local authorities take to prediction in order to estimate and curb pupil drop-out.

Let me, for an empirical overview of predictive practices, refer the reader to a recent report by AlgorithmWatch, a German watchdog organization. Covering 12 European countries, they document the discussions, regulatory proposals, and oversight mechanisms pertaining to algorithmic decision-making, as well as the systems of the kind actually in use in both private and public organizations (AlgorithmWatch 2019). For our purposes, their sections about algorithmic decision-making in action—which is a more technical term for what I refer to as predictive practices—are instructive, since the rapporteurs focus explicitly on 'systems that affect justice, equality, participation, and public welfare' (AlgorithmWatch 2019, p. 9). That focus is quite in line with the focus of my research.

As can be glanced from this overview, many of the institutions that rely on prediction are characterized by relations of power. Principals lord it over their agents, resorting to disciplining in case of malfunctioning. Otherwise, dependencies are involved which carry obligations that have to be met. Or agents wield power in the name of state or local authorities. Thus always a normative order of a kind reigns—rules and standards of behaviour have to be met. The exception to this rule is data base marketing in the corporate sector: the efforts to nudge or win over customers are inspired by norms, but these are not dictated to clients as an absolute command. The same goes for some prediction-based campaigns by governmental agencies that deserve the epithet "paternalistic" (say, to convince people to quit smoking or collect rent subsidies they are entitled to).

So this is my first rejoinder to the theories explicated above. Profiling leads to social sorting, indeed. But more often than not such sorting categories are the *expression* of an underlying normative order. Appearing to deviate from it in the (near) future leads to close attention—its form depending on the specific context. Once individuals emerge from ML as suspects, they are seen as likely to transgress the normative order. Accordingly, they may obtain a loan from credit agencies on restrictive conditions only, are taken out of the queue for questioning by airport officials, are placed under increased police supervision, and so on. This close attention is a *precursor* of possible sanctions. In some cases, though, the individual is considered too big a risk already; then the suspicion alone is reason enough to issue a sanction *immediately:* the loan is refused, the plane passenger is sent home straight away, the prisoner is not released on parole, and the like.

Already, we see emerging one of the central elements of a Foucauldian regime: normation. ML does not *create* suspect categories as is often maintained. Instead, the exercise takes as its starting point an apt "target variable" that is supplied by the institution. In the institutional contexts we are dealing with here, those are the subjectivities that are the focus of scrutiny for the personnel involved: the good (or bad) credit risk, the model plane passenger, the obliging worker, etc. In case of airline passengers, the construction of the binary scale is rather easy: does the passenger intend to blow up the airplane or not? But it can also be hard work—often referred to as the "art" of data mining—to construct a target variable that adequately expresses the concerns of the institution involved. A standard example is creditworthiness: what is to count as such? In the past banks have had tough discussions on how to operationalize the concept and divide it into adequate class variables (Barocas and Selbst 2016, pp. 678–680).

The ML effort then starts to work on the basis of this particular target variable (so called "supervised learning"). Data are used to train models; the data used are referred to

as "training data". These should contain enough data points of all categories of the target variable ("class variables")—otherwise the exercise is futile. The aim is to create a model that separates the class variables from each other. Referring to the examples above: separating the good credit risks from the bad ones, the terrorist passengers from those who have no such intentions, and so on. The model is geared to separate the deviants from the conformers.

But where do these data come from? Obviously, the institutions involved collect data about their subjects or clients, as they already did long before the era of Big Data began. But nowadays additional datasets are procured from outside. Data are collected from everywhere, with every step we take (especially on the Internet) we create a digital footprint. Our behaviour in other institutions, our search and shopping behaviour, our social media presence, all gets recorded. Data that sometimes have to be provided, sometimes have been provided voluntarily. In the latter case, we are not even aware that such data are being stored. Now, all such acquired datasets are routinely exchanged *between* institutions (without much regard for the law) (cf. Wigan and Clarke 2013). Data-brokers even make a living from the practice. So modern ML operates on an elaborate pool of datasets that have been gathered in a variety of contexts.

Can this be interpreted using the metaphor of the Panopticon? I think this is indeed the case, in a double sense. As far as the "focal" institution is concerned, its subjects must assume that relevant data are generated and collected all the time—a routine Panopticon. But in addition, they must assume that in many other Panoptica in which they are entangled, other digital traces about them are monitored and stored. Subsequently these data may be imported backwards into the focal institution that we are considering. All such imports boost ML efforts considerably, often in surprising ways. Thus the panoptic gazes of many different contexts are coupled together: a "*Polypanopticon*". Many hitherto separated domains of life become intertwined.

Two comparisons suggest themselves. For one thing, predictive practices are executed in a true "surveillant assemblage"—the convergence of once discrete surveillance systems (a term coined by Haggerty and Ericson 2000). For another, concerning the plurality of datasets involved, note the parallel with the metaphoric concept of "rhizomatic" growth, as coined by Deleuze long ago (cf. discussion in Haggerty and Ericson 2000). You may uproot one plant (read: dataset), but the damage is already done, the roots have branched underground (read: multiplied) so that other exemplars of the plant will emerge in abundance (or already have). Once a dataset has been created, its multiplication is unstoppable.

The importance of this data coupling relates directly to the character of ML. Predictive modelling can only benefit from bringing in as many datasets as possible about the persons involved. The more variety in independent variables the better the outcomes (though with a caveat: experts warn that extra datasets can also import a lot of noise that effectively supresses the signal one is looking for). As has been described above, ML has no conception of causality—all data are taken as input. As a result, surprising correlations may be produced. Compare the terrorists-to-be who are halted for inspection—because they paid in cash and/or carried no luggage; or the tax payers whose tax declarations are selected for auditing—because they contributed a lot of money to charity. Thus, precisely this abundant coupling of far-away datasets about individuals may boost the results of ML.

So taken together, predictive modelling may be interpreted as a case of Foucauldian discipline, though with a twist. As far as normation is concerned, prediction represents an additional mechanism which *extends* a normation already in existence. Compliance to the norm is measured in in enhanced predictive fashion which subsumes the classical way of measuring. As far as the data are concerned, these are procured by both the focal institution and by other contexts (the "Polypanopticon"). An institution can glance sideways as it were, and peer into the data pools of other institutions. The discipline of ML thus strengthens existing relations of power by harnessing the predictive power of Big Data. Just as Foucauldian discipline gradually crept into the various institutions and strengthened their modalities of power a few centuries ago, so do predictive practices at the present day.

Note the curious conception of time and timing of this strengthened discipline. Primary discipline is put into action as soon as a subject breaks—or appears to be about to break—the norm; the secondary mechanism of discipline based on prediction as an outcome of ML may suggest an intervention even before that: as soon as a subject appears likely to be breaking a norm sometime in the future. The former discipline had to wait until norm deviance happens or seems immanent; the latter anticipates it and acts accordingly. In the case of burglary: wait until the break-in actually happens, or appears to be about to happen (primary discipline), or in anticipation put the suspect under close surveillance (secondary discipline).

For the decision subjects this extended disciplining has grave consequences (to which I come back more fully later). Algorithmic decision-making changes the way in which organizations approach their subjects. In a nutshell: with primary disciplining a decision was based on causally related variables or reputational indicators. Moreover, a decision could in principle be explained to the decision subject. Consequently, they could try to adapt their 'parameters' in order to do better. As well, the subject could contest a decision in case of a disagreement. Secondary disciplining changes all that: decisions are based on any number of associated variables as long as they add to the accuracy of the algorithm.

Further, they cannot be explained in any way since algorithms are (usually) opaque. As a result, the subject has no way of understanding a decision, and, as a corollary, cannot contest it—grounds for decisions are simply lacking.

Individualized risk

Are ML models geared to governance by risk? Are individuals reduced to risk categories that have to be managed? I think this conception referred to above has become outdated. In former days ML indeed produced so-called profiles of high-risk deviants. As soon as individual characteristics matched such a profile, they were considered to represent high risk. Such profiles steered the interventions of corporations (such as insurance companies and banks) and public bodies (such as security forces).

Nowadays most ML models no longer have profiles as their output. Profiles are a special kind of decision tree (one-sided) that is interpretable. But decision trees and classifications in general do not have easily interpretable outcomes in terms of the underlying variables. Nor do neural networks allow any such explanation. The models have effectively become black boxes. But on the plus side, they yield more precise outcomes; and the predictive probabilities of being a deviant are tailored to each individual. So governance nowadays has receded from governing risk groups; governance by *individualized* risk has taken over.

Critics might object that the subject is still not addressed as such, but as the risk (s)he represents. But that has always been the case in the contexts we are considering here: the bank or the police were always on the lookout for suspicious signs that might indicate someone was out of line. ML just produces a new—more powerful—indicator of suspicion that is added to the repertoire of the authorities concerned.

Is the individual just dissolved into his/her constituent data—never to emerge again? On the contrary, I would say. Data about individuals are used in two different ways. For one thing, a set of them are used for training purposes. These should be data that are reliable and vetted; especially concerning the target variable—otherwise the ML effort is useless. In a way, these "divided" individuals have fulfilled their purpose and will not reappear. For another, once a model has been produced it stands ready for predictive purposes. That is, new individuals who appear on the scene for consideration are subdivided (again) and their data fed into the algorithm. As output, say, a risk score appears. That score is influential in the decision-making of the institution about that individual. It will be a decision specifically geared to that individual—instead of disappearing, he/she has resurfaced.

Another contention referred to above was, that governance by risk is no longer interested in drawing the line between normal/abnormal people. This is a curious assertion that completely disregards what is at stake with ML. The starting point is always an adequately defined "target variable". If it is a binary one, the variable reads good credit risk/bad credit risk; terrorist inclinations no/yes; or more generally: normal/abnormal. The modelling then sets out to draw the best line between the normal ones and the abnormal ones, to sort the normal from the abnormal training instances. So predictive modelling is precisely geared towards separating normal and deviant people—that is what the exercise is all about. And note: in the old days of profiling, as well, such a separation was always the ultimate goal.

Matzner's performance view on data

This Foucauldian view on Big Data and discipline is actually quite close to the view of Tobias Matzner, who recently published several articles about Big Data. He explicitly chooses to narrow down the usual (broad) definition of surveillance to data gathering in contexts of discerning "suspects" (Matzner 2016, p. 200), roughly those of security and police. With that kind of surveillance, he writes, 'subjectivizing moments happen' (Matzner 2017, p. 31). In particular, data are used to identify 'suspects' and discipline them if necessary (Matzner 2017, p. 44). Properly speaking, they are established as "criminals before the act" (a hint to Foucault that I come back to later; Matzner 2017, p. 39ff.) He develops this Foucauldian view as a counterpoint to the views of Amoore, Rouvroy, and others.

Nevertheless, I have some minor points of criticism of his stance. For one thing, he advocates a "performance view" on data (as opposed to the usual "representationalist view" of data: how accurate are the data?) (Matzner 2016). However, as is clear from my description of ML, it is not data that perform but the *algorithms* that are produced from them. The neural networks and decision trees create the suspect, not the data. And depending on how well the algorithms are crafted, they may perform quite differently. In line with this, we talk about governance by algorithm; no one has ever coined the term governance by data.

For another, the creation of suspects also happens in other contexts than security alone. Compare above: suspicions of tax fraud, insurance fraud, money laundering, etc. A *plurality* of suspects can be created. Finally, in Matzner's work (especially Matzner 2017) he often refers to the phrase: creation of subjectivities. As well, he speaks about norms that are derived from data. This creates the impression that such subjectivities turn up as an outcome of data processing—there were none before. But as elucidated above, ML (at least supervised learning)

always starts from a particular subjectivity salient to the institution. ML methods then aim to develop an algorithm that reproduces this (primary) normation with the highest possible accuracy.

Foucault: prescient about prediction?

Many an author in the surveillance literature I referred to (Amoore and Matzner in particular) alludes to specific sections in Foucauldian texts that are interpreted as *precursors* of the procedure of prediction which is executed by ML. It seems worthwhile to bring up these passages here. Not for the sake of a discussion of what Foucault had or had not foreseen, writing in the period around the 1970s before the age of Big Data dawned. These texts can help, though, to clarify what modern predictive methods have achieved compared to the tools of the past. In comparison, I would argue, they embody conceptual steps that are by no means trivial.

The first reference to be treated here is a section of *Discipline and Punish*, in which Foucault gives an account of the changes in the French penitentiary system during the nineteenth century: the "delinquent" substitutes for the "criminal" (Foucault 1977, pp. 251–254). Once the convict enters the prison, his criminal act is no longer important, but his life is. The causes of his crime are to be found in a biographical investigation, which delves into 'psychology, social position and upbringing' (Foucault 1977, p. 252). Thus 'it establishes the "criminal" as existing before the crime and even outside it' (Foucault 1977, p. 252) (in the French original: 'Il fait exister le "criminel" avant le crime et, dans la limite, en dehors de lui'; Foucault 1975, pp. 255, 256). A whole typology of delinquents was being developed; each type had to be treated by a specific prison regime. Thus, the penitentiary system took over from the juridical system. In the new system, the delinquent is to be observed permanently. Do the correctional measures succeed in reforming them?

Actually, around the same time that Foucault wrote *Discipline and Punish*, he delivered a series of lectures at the Collège de France called *Abnormal* (Foucault 2003). In the lecture of 8 January 1975 this theme of the "delinquent" was rendered even more forcefully (Foucault 2003, pp. 16–25). In particular, the role of psychiatric experts obtained more emphasis. On the court's request they seek to explain how the crime has come about, *assuming that the suspect did indeed commit it* (my italics) (Foucault 2003, p. 17). Thus they create a 'psychologico-ethical double of the offense' (Foucault 2003, p. 16). Then, when the judge has to pass judgment on the suspect, (s)he will not do so on the basis of the crime committed, but of the forms of deviant conduct that are the substance of the constructed double. The aim of such expert opinion is, remarks Foucault again,' to show how the individual already resembles his crime before he has committed it' (Foucault 2003, p. 19). Subsequently, the delinquent becomes 'the object of a technology and knowledge of rectification, readaptation, reinsertion, and correction' (Foucault 2003, p. 21). So the psychiatrist becomes a psychiatrist-judge (Foucault 2003, p. 23), greatly influencing the question of guilt and the ensuing follow-up measures of correction.

Both Amoore, citing *Abnormal* (Amoore 2013, p. 49), and Matzner, citing *Discipline and Punish* (Matzner 2017, pp. 38, 39), refer to passages of this kind. They are struck especially by the—indeed striking—phrase that 'the individual already resembles his crime before he has committed it'. They interpret Foucault's assertions as follows. The experts were actually involved in finding the biographical traits that caused the crime to occur. In the process they used the (primitive) theories of social conduct that were beginning to evolve at that time. From that point onwards, it is 'a rather small step' (says Matzner 2017, p. 40) to reverse the time arrow and initiate the process of predicting such crimes; the essence of modern-day surveillance systems. Of course, modern approaches seem to lie around the corner. But the differences between the prediction associated with delinquency and modern-day ML are considerable; several conceptual steps have to be taken in between.

For one thing, theories of social and psychological causation were more primitive. But apart from that, Foucault's experts were reasoning *backwards*, on the assumption that the accused was guilty; then it is always easy to find contributing factors. And as Foucault forcefully remarks (Foucault 2003, pp. 22, 23): in case of doubt, those with a monstrous personality were most usually considered the guilty ones. So much for deriving guilt from the accused's past. More daunting is the task in reverse: predicting who is guilty based on a set of biographies. Moreover, the experts in Foucault's court room reasoned with theories characterized by causation—which factors caused the crime to occur? Modern day ML goes beyond causation: all variables pertaining to individuals are potentially taken into consideration, ML is associational.

Finally, modelling by ML tries to separate the guilty from the innocent. To that effect ML needs verified data for training about the guilty *and* the innocent. Biographies of both categories are needed. A vexing problem usually presents itself: data about the former are rarer than data about the latter (class imbalance problem). That problem then has to be remedied before modelling starts (mainly by under- or oversampling, see above).

Another passage from Foucault, this time from his 1978 lectures (Foucault 2007), also deserves consideration. Working on a perceived shift in governance from the individual to the group, the population, in lecture 3 of 25 January 1978 he engages in a discussion of the governmental approach

to small pox epidemics in the eighteenth century (Foucault 2007, pp. 85–91). In that approach, Foucault observes, the usual process of *normation* (in which a norm is established and subsequently enforced by disciplining the individual) is no longer at work. Instead, Foucault argues, we witness a more complicated approach that he dubs *normalization*. Data are collected about the incidence of the disease amongst different subpopulations (older vs. younger people, a particular milieu or profession, town vs. countryside, etc.). These are, say, normal distributions, each with their own mean value. From such data one may derive which populations are most at risk. In this case of smallpox that Foucault is discussing: new-born children were found to be especially vulnerable (a chance of 2 in 3 of catching it). Targeted intervention then strives to bring this highest mean back in line with the overall mean (for smallpox: 1 in 8). The intervention consists of inoculation.

Matzner uses this passage to emphasize how "data-driven" surveillance already was (Matzner 2017, pp. 37, 38). I highlight the passage for quite another reason: it shows the way, so to speak, to governance by risk, away from individual disciplining. Society's scarce resources are best spent with a cool-headed risk approach—instead of wanting to cure everybody. Compare the earlier approach to smallpox (as described by Foucault): isolate the sick from the healthy and subsequently try to cure them all. I presume that the theorists mentioned in the above have taken their lead from these preoccupations of Foucault when they coined their theories about the (supposed) shift from governance by discipline to governance by risk.

But again, this "normalization" approach is unlike the modern day ML-based prediction—although it comes much closer than the preceding approach of medical experts in the courtroom. Using statistical techniques it can identify high-risk groups, allowing targeted governance. In a way, a profile is established of the high-risk patients. With ML, in comparison, the procedure of generating a model from training data is quite specific, and completely different from classical statistics methods. Moreover, ML takes all associations into account, not only the causal ones. As a corollary, it can identify all kinds of groups, both "natural" groups and groups artificially created for the purpose. Finally, ML delivers tailored scores for each individual.

Individual resistance

So modern day prediction as executed by ML again centres on the individual. It has evolved into a truly disciplinary apparatus. Foucauldian discipline was supposed to exercise a self-disciplining force on the inmates. What about the reactions of those subjected to predictive discipline: do they comply, negotiate, or resist (cf. Lyon 2003, p. 20)? What are their options? And does by any chance resistance in organized fashion arise?

Many of those subjected to predictive modelling will not be aware of it. Only when a decision affects them, based on one model or another, might awareness arise. Most probably they will ask for an explanation of how the decision came about. Unfortunately, such an explanation is hardly ever given. And when details *are* provided, they tend to be uninformative. Legal clauses to that effect are in operation, but have little effect (cf. Zarsky 2013, p. 1510ff., p. 1523ff. for the American situation). So basically, all details about decision-making based on predictive algorithms remain opaque.

This darkness is exacerbated if one starts thinking about what kind of personal data have entered into the equation. Obviously, data about one's relation to the institution concerned (such as the bank, police, or airport security) may be involved. Apart from these data that have been handed over consciously, however, transactional data about one on the Internet may be involved as well. One leaves electronic footprints everywhere. Usually one does not think about this for a second, let alone about the possible consequences for predictive modelling. Daniel Susser (2016, pp. 224, 225)—employing a notion developed by Erving Goffman—refers to this information as "given off," as opposed to the usual information being given.

Now this predicament—an algorithmic decision without explanation and hardly a clue about the kind of data that have been fed into the modelling—can properly be described as a Kafkaesque situation (Susser 2016, pp. 231, 232, borrowing the insight from Solove). One is a suspect, but what is the charge? The answer will not come. The disciplining involved is diffuse. In Foucauldian discipline, which very much centred on the body, it was abundantly clear why you failed and what you should do next in order to succeed—compare a soldier who failed a shooting exercise and knew immediately that he was supposed to do it once more. With predictive disciplining, centred on the mind, you only know that you have apparently already been failing—but why and how to do better is in the dark.

The only thing left is to use your imagination and make a list of all the kinds of possibly suspicious behaviours that may have been picked up by the modelling. Do you want a loan? Avoid the use of a credit card, and do not go to the casino. Do you want to take the airplane? Do not buy the ticket with cash, and do not only buy a one-way ticket. Do you want to overhaul your garden? Do not order large quantities of fertilizer over the Internet. So ultimately one goes into a mode of self-censorship, fuelled only by one's imagination, in order to avoid supposedly suspicious indicators and thereby escape intensified scrutiny by the institutions involved. Our lives are forced into conformity.

Organized resistance

Organized resistance potentially offers some respite from this Kafkaesque predicament. At several levels, various groups on various continents are calling for more transparency as far as predictive modelling is concerned. Not only academic circles and privacy activists (EPIC), but also computer professionals (ACM, IEEE), standard setting bodies, and various governmental bodies and parliaments in the EU and the US have issued calls of the kind. Such transparency is intended to contribute to restoring accountability for computerized systems that assist humans in their decision-making.

It is crucial, as far as transparency is concerned, to distinguish its possible beneficiaries: intermediate oversight bodies, or the public at large (cf. de Laat 2018, p. 527). Accordingly, transparency can be enlarged in steps, each step enlarging the circle of those in the know. The first option to consider is installing *oversight bodies*. Think of governmental agencies, or external bodies of experts. It is their task to certify that professional standards of accuracy, fairness, robustness and the like are adhered to all along. Two varieties of such oversight can be distinguished. With white-box transparency, experts obtain full insight into the datasets used, the process of modelling and its algorithmic outcome, and how the final model is used. With black-box transparency the same applies, except that the algorithm in use remains opaque and can only be tested from outside the organization.

What does this auditing bring to us ordinary citizens? At least we may be ensured that the modelling that pertains to us is adequate and up to professional standards. We may still be in the dark as to what led to a decision, the fog from Kafka's *'Der Prozeß'* persists, but at least we have a consolation in its non-arbitrariness.

The most promising option would be, of course, if transparency to *the public at large* could be achieved. While disclosing the very datasets used in modelling would not seem a good idea (in view of privacy considerations), having out in the open the models in use and how they are applied to concrete cases would be a good idea from the point of view of ordinary citizens who demand to be informed of the reasons for algorithmically-inspired decisions. At long last we would be able to see the reasoning behind predictions that affect us.

These hopes for an understanding of the algorithms may, however, turn out to be just a fata morgana. More often than not, ML models are intrinsically opaque (what follows is excerpted from de Laat 2018, para. 7). While simple classifiers are interpretable, modern classifiers use multiple trees in their construction, hundreds of them. Outcomes thus can no longer be interpreted easily. Furthermore, neural networks have always been inscrutable—by design. The main trend in ML is towards greater efficiency of outcomes; interpretability is relegated to the background.

Only in the rare case that a model is interpretable and can be 'read' easily (like a one-sided decision tree or a score card), would we be able to glimpse the main factors that count. As a corollary, subjects would be able to know why an organization affected them through a particular decision. But at the same time, paradoxically, options for resistance are opening up. The system can be 'gamed': detect the proxies involved and try henceforth to evade them. The feasibility of this, though, depends on the type of proxies involved (the following section about proxies is based on de Laat 2018, para. 5).

When proxies refer to specific behaviours, evasion seems obvious. A tax evader learns from profiles for tax evasion that high donations to charity are a red flag; in future he/she cancels these donations. In the search for a loan one discovers that using a credit card is a bad omen; so one no longer uses it. Potential terrorists learn that wearing a Palestinian scarf sends a clear signal; so they learn to avoid the scarf. Less obviously, they may learn that paying in cash, and for a one-way plane ticket at that, is considered suspicious; so they obtain a ticket in another way. However, when the proxies refer to personal characteristics, evasion becomes harder. Being an accountant turns out to arouse suspicions of evading taxes; being a Muslim of Middle-Eastern appearance cannot but arouse fears of terrorist intentions. Such characteristics can hardly be evaded (or only at a price).

Unsurprisingly, private organizations, and in their wake, public organizations, are dead set against such total transparency. The prediction models in use are considered to be their intellectual property. The models give a competitive advantage that they are not prepared to give up easily. So they guard their algorithms as a trade secret, embed watermarks in them to prevent theft, or apply for patents on the underlying methods (which strengthens protection even more). Moreover, as far as interpretable ML models are involved, the prospect of subjects gaming their models would necessitate additional modelling efforts. Either the models have to be made more robust against gaming, or the gameable proxies have to be omitted from the modelling (which might well delete valuable information).

So, for the near future, limited transparency—for oversight bodies only—seems to be the only feasible option. Such a development would only mildly soften our Kafkaesque predicament. We can be sure that the predictions that concern us are up to the professional standards of modern ML; their accuracy, fairness, and the like are beyond doubt. In some contexts you will turn out to be a suspect—time to ponder if you already are one after all?

Conclusions

In many contexts, predictive practices based on ML amount to disciplining. Norms are to be respected; a prediction that deviance is around the corner, is enough to trigger close attention, restrictions, or even sanctions. Governance by disciplining has definitely not given way to governance by group risk. In reaching this conclusion it has been shown that tools from the Foucauldian toolbox can still usefully be deployed for an analysis of present day surveillance. Notice, though, that Foucauldian discipline is specific and focused on the body—while predictive discipline is more diffuse and focused on the mind.

For those subjected to predictive disciplining, transparency of the algorithmic process to oversight bodies would seem the maximally attainable option in the near future. We may rest assured that the discipline affecting us conforms to the best practices of the ML community. In rather utopian fashion, though, I can imagine two alternative future scenarios that would give some relief from Kafkaesque gloom.

First, recently, trimmed-down varieties have been suggested of full transparency to the public at large. The reasoning behind a decision about a particular data subject can be derived and communicated to him/her upon request ("subject-centric explanations"; cf. algo:aware 2018, p. 25). Some examples are readily mentioned. A sensitivity analysis (aka counterfactual explanations) is intended to answer the question of how much a focal subject's input data have to change in order to change the outcome. The quantitative input influence approach measures the influence of various inputs on decisional outcomes for a particular data subject. In quite another approach, new models that are easily interpretable are learned locally, in an area around the focal subject; they mimic the local behaviour of the full model (a technique called LIME). Such local transparency, of whatever flavour, may have more chances to be realized in practice than full transparency, since these solutions do not require the full algorithm to be disclosed to anyone. Therefore, the objections against disclosing intellectual property and/or of creating an invitation to game the system carry less weight.

Secondly, in a broader time frame we may draw courage from a movement in ML that brings the interpretability of models into the foreground. In particular, only *causal* variables are to be entered into the modelling. Especially for medical contexts they argue that uninterpretable models are simply no longer acceptable, neither to medical personnel nor to patients. The challenge is, whether accuracy of modelling can be salvaged all along. If this demand ever reaches the disciplinary contexts discussed above, it would mean that those in authority can no longer hide behind opacity and remain silent. Instead, they are held to account and forced to open up their models to the public and provide reasons for decisions upon request. Obviously, complications related to gaming the system would have to be sorted out.

As a final reflection, we may take an even bigger step forward (or backwards?) and argue that this predictive disciplining before-the-deed should be rejected altogether. It is this conclusion that Matzner draws from his investigations of the subjectivities created by predictive techniques. However useful it may be to open up the black boxes of algorithms in use, we should, he argues, move beyond the issue of transparency and ask ourselves whether we want them to 'determine our lives' at all (Matzner 2017, p. 44).

References

algo:aware (2018) Raising awareness on algorithms. Procured by the European Commission's Directorate-General for Communications Networks, Content and Technology. State-of-the-Art Report/Algorithmic decision-making. Version 1.0, December 2018. www.algoaware.eu/state-of-the-art-report/. Accessed 20 July 2019.

AlgorithmWatch (2019) Automating society: Taking stock of automated decision-making in the EU. A report by AlgorithmWatch in cooperation with the Bertelsmann Stiftung, supported by the Open Society Foundations. 1st edition, January 2019. www.algorithmwatch.org/en/automating-society. Accessed 20 July 2019.

Amoore, L. (2011). Data derivatives: On the emergence of a security risk calculus for our times. *Theory, Culture & Society, 28*(6), 24–43.

Amoore, L. (2013). *The politics of possibility: Risk and security beyond probability*. Durham, London: Duke University Press.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104,* 671–732.

de Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? *Philosophy & Technology*, *31*(4), 525–541.

Deleuze, G. (1992). Postscript on the societies of control. *October, 59*(winter), 3–7.

Erwin, S. (2015). Living by algorithm: Smart surveillance and the society of control. *Humanities and Technology Review, 34*(Fall), 28–69.

Foucault, M. (1977). *Discipline and Punish: Birth of the Prison* (Alan Sheridan translator). New York: Vintage Books. Translation of *Surveiller et punir: Naissance de la prison*, Paris: Gallimard, 1975.

Foucault, M. (2003). *Abnormal*. Lectures at the Collège de France 1974–1975. London: Verso. Translation of *Les Anormeaux*, Paris: Gallimard, 1999.

Foucault, M. (2007). *Security, Territory, Population*. Lectures at the Collège de France 1977–1978. London: Palgrave Macmillan. Originally published in French in 2004.

Haggerty, K. D. (2006). Tear down the walls: On demolishing the panopticon. In D. Lyon (Ed.), *Theorizing surveillance: The panopticon and beyond* (pp. 23–45). Cullompton: Willan Publishing.

Haggerty, K. D., & Ericson, R. V. (2000). The surveillant assemblage. *British Journal of Sociology, 51*(4), 605–622.

Lyon, D. (2003). Surveillance as social sorting: Computer codes and mobile bodies. In D. Lyon (Ed.), *Surveillance as social sorting: Privacy, risk, & digital discrimination* (pp. 13–28). London: Routledge.

Lyon, D. (2006). The search for surveillance theories. In D. Lyon (Ed.), *Theorizing surveillance: The Panopticon and beyond* (pp. 3–20). Cullompton: Willan Publishing.

Matzner, T. (2016). Beyond data as representation: The performativity of big data in surveillance. *Surveillance & Society, 14*(2), 197–210.

Matzner, T. (2017). Opening black boxes is not enough: Data-based surveillance in *Discipline and Punish* and Today. *Foucault Studies, 23,* 27–45.

Rouvroy, A. (2012). The end(s) of critique: data-behaviourism vs. due-process. In M. Hildebrandt & E. De Vries (Eds.), *Privacy, due process and the computational turn. Philosophers of law meet philosophers of technology (Chap. 5)*. London: Routledge.

Susser, D. (2016). Information privacy and social self-authorship. *Techné: Research in Philosophy and Technology, 20*(3), 216–239.

Valverde, M., & Mopas, M. (2004). Insecurity and the dream of targeted governance. In W. Larner & W. Walters (Eds.), *Global governmentality: Governing international spaces* (pp. 233–250). London: Routledge.

Wigan, M. R., & Clarke, R. (2013). Big data's big unintended consequences. *Computer, 46*(6), 46–53.

Zarsky, T. Z. (2013). Transparent Predictions. *University of Illinois Law Review, 2013*(4), 1503–1570.